Supplementary information: Challenges to aboveground biomass prediction from waveform lidar

Jamis M. Bruening[1], Rico Fischer[2], Friedrich J. Bohn[2], John Armston[1], Amanda H. Armstrong[3,4], Nikolai Knapp[2], Hao Tang[1,5], Andreas Huth[2,6,7], Ralph Dubayah[1]

[1] Department of Geographical Sciences, University of Maryland, College Park, MD 20740, USA

[2] Department of Ecological Modeling, Helmholtz Centre for Environmental Research (UFZ), 04318 Leipzig, Germany

[3] Department of Environmental Sciences, University of Virginia, Clark Hall, Charlottesville, VA 22902, USA

[4] Universities Space Research Association, Goddard Earth Sciences Technology and Research Studies and Investigations, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA

[5] Department of Geography, National University of Singapore, Kent Ridge, Singapore 117570

[6] German Centre for Integrative Biodiversity Research (iDiv), 04103 Halle-Leipzig-Jena, Germany

[7] Institute of Environmental Systems Research, University Osnabrck, 49076 Osnabrck, Germany

E-mail: `jamis@umd.edu`

1. FORMIND calibration

We calibrated a regional version of FORMIND to represent the tree geometry and stand structure for Northeastern US forests using data from the United States Forest Service Forest Inventory and Analysis (FIA) program, accessed with the rFIA software package [1]. Using FIA data we segmented the region's 27 most abundant tree species (based on total species basal area) into 9 different plant functional types (PFTs) (table S1). Together these 27 species account for upwards of 85% the region's total basal area [2]. We calibrated the suite of allometric equations required by FORMIND for tree construction for each PFT using FIA measurements, and we calibrated tree biomass as a function of diameter according to the national-scale allometric equations from Jenkins *et al.* [3]. As the FF simulates forest structure, not growth, of trees, we only needed to calibrate physical tree geometry equations, which are straightforward. Calibrations of growth rates, photosynthesis, and other dynamic processes are more involved, but were not necessary for this analysis, and ignored.

2. Forest Factory

The Forest Factory (FF) simulates unique, 20 m x 20 m stem maps, and the structure of these stem maps is then generated by FORMIND in the same manner as the *in situ* stem maps. The purpose of the FF is to simulate the diversity in forest structure within a region, based on varying the PFT compositions and stem-size distributions across a large number of 20 m x 20 m simulated forest stands. During the construction of a forest stand, the FF iteratively adds trees to the plot by sampling from one of several predefined stem size distributions, and assigns each tree to a PFT. With the addition of each tree the FF checks to make sure there is space and resources (light, water, etc.) available to accommodate the new tree. In this way the FF ensures that stands are not unrealistically crowded with trees, and could theoretically exist in reality. For more on the Forest Factory see Bohn and Huth [4].

The FF outputs 100 simulated stands at once, arranged in a 4 ha, 10 x 10 grid of 20 m x 20 m stands. While each stand is generated independently from the others, it is output directly next to other stands in the simulation space, and thus the crowns of trees near the edges of each stand extend over the plot boundaries into the adjacent stands. To avoid edge effects that would otherwise alter the waveform shape of each stand, we implemented the following process for every simulated stand in the FF database. We generated each FF stand's LiDAR point cloud with periodic boundaries, meaning that the tree crowns which extend over a plot boundary enter back into the plot on the opposite edge and do not enter into the adjacent stands, following the procedure of Knapp *et al.* [5]. This point cloud was then replicated into a 3 x 3 grid, so that when positioned directly to each other the periodic boundaries align and the nine 20 m x 20 m point clouds represent a homogeneous 60 m x 60 m point cloud with structural characteristics identical to the original 20 m x 20 m stand. The GEDI waveforms were then simulated in the center of this expanded point cloud, so that no tree crown influenced the waveform that was not accounted for in the stem map.

Table S1: The 27 most abundant tree species within the Northeast US are grouped into nine PFTs based on common structural characteristics and shade tolerances. Maximum height and diameter values were derived by assessing the distributions of each variable for each PFT using all the most recent FIA data for the Northeast USA. N indicates the number of samples used for each PFT to calibrate the tree geometry equations needed in FORMIND [6].

PFT	**Name**	**Scientific**	**Common**	**Tolerance**	**Maximum height**	**Maximum diameter**	**N**
1	fir-cedar	Abies balsamea Thuja occidentalis Picea mariana	balsam fir northern white-cedar black spruce	tolerant	35	0.87	71544
2	hemlock-spruce	Tsuga canadensis Picea rubens Picea glauca Picea abies	eastern hemlock red spruce white spruce Norway spruce	tolerant	43	1.11	67419
3	pine	Pinus strobus	eastern white pine	intermediate	54	1.60	21362
4	maple-beech	Acer saccharum Fagus grandifolia Tilia americana	sugar maple American beech American basswood	tolerant	45	1.34	56484
5	maple-ash	Acer rubrum Fraxinus americana Prunus serotina Fraxinus pennsylvanica	red maple white ash black cherry green ash	intermediate	45	1.13	66199
6	oak	Quercus rubra Quercus alba Quercus velutina Quercus prinus	northern red oak white oak black oak chestnut oak	intermediate	46	1.57	12243
7	birch	Betula alleghaniensis	yellow birch	intermediate	40	1.28	19362
8	birch-aspen	Betula papyrifera Populus tremuloides Betula lenta Populus grandidentata	paper birch quaking aspen sweet birch bigtooth aspen	intolerant	39	0.93	26622
9	understory	Acer pensylvanicum Ostrya virginiana Acer spicatum	striped maple eastern hophornbeam mountain maple	tolerant	29	0.50	5023

3. OLS method and results

We did not test all the possible combinations of model equations that could be used to predict AGBD from waveform RH metrics in 10m increments (RH10 - RH90 and RH98). Instead, we developed a set of 18 candidate models; each model contained one of three different upper-canopy metrics (RH98, RH90, RH80) and one of three different mid-canopy metrics (RH50, RH40, RH30), as the combination of upper and lower canopy metrics has been effective in modeling AGBD from GEDI waveforms [7, 8]. The first nine models were all the possible combinations between these groups, and the second nine models were the same as the previous and also included the interaction term between the upper- and mid-canopy variables. Performance of the top five candidate OLS models was similar when tested on the 50% of FF stands that were not used in model training, and final model selection (figure S1) was based on the lowest nRMSE value (table S2). Once the final model was selected, we performed the back transformation on the fitted values and 90% prediction interval to obtain the predicted values and prediction interval in meaningful units (Mg ha^{-1}).

Table S2: Top five OLS models, sorted by nRMSE. In the formulas, y is the response variable, AGBD. The reported nRMSE and R^2 were calculated from the back-transformed predictions. Model 15 was selected as the final model.

Model	**nRMSE (%)**	**R^2**	**Formula**
15	32.08	0.829	$\sqrt{y}$ ~rh98 + rh40 + rh98*rh40
12	32.19	0.828	$\sqrt{y}$ ~rh98 + rh50 + rh98*rh50
14	32.47	0.825	$\sqrt{y}$ ~rh90 + rh40 + rh90*rh40
11	32.56	0.825	$\sqrt{y}$ ~rh90 + rh50 + rh90*rh50
2	32.59	0.824	$\sqrt{y}$ ~rh90 + rh50

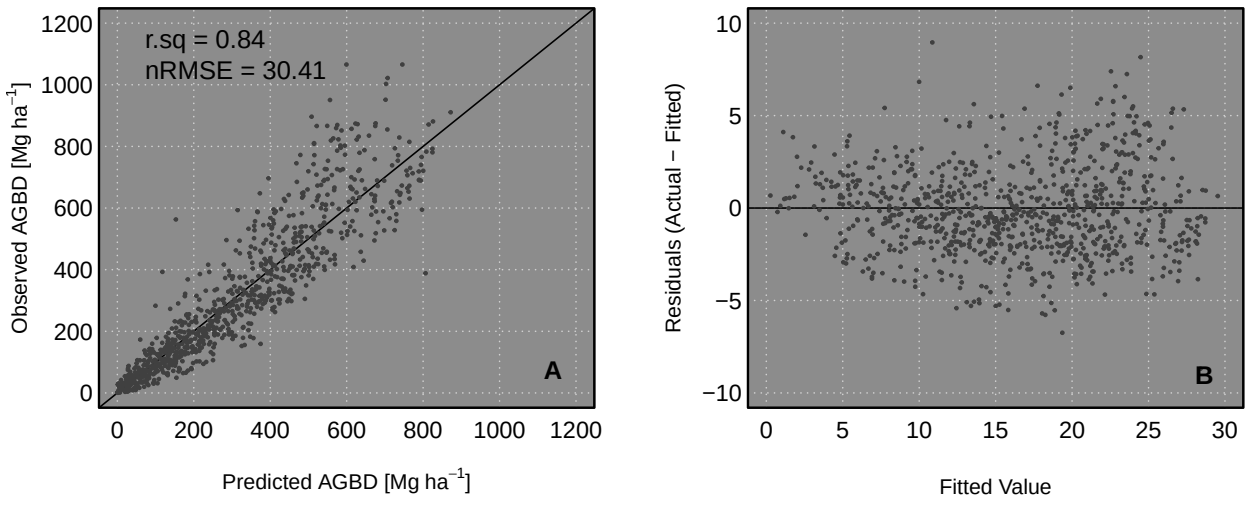

Figure S1: The final OLS model applied to one of the 500 testing subsets (A), performs well and satisfies the assumption of constant variance. The residuals plot (B) is in square-root transformed units.

4. WFM method and results

Forty eight stem map waveforms did not have at least 100 FF waveforms with an $r > 0.75$. The relative overlap algorithm is highly sensitive to slight differences in waveform shape, and we intentionally set this strict criteria (100 matches with $r > 0.75$) to ensure we only analyzed stands with 100 similar waveforms. Waveform shape is also highly sensitive to slight differences in forest structure within a plot, and it was our expectation that all 428 plots may not have 100 matches. We tested various r thresholds and the number of matches required for each *in situ* stem map, and arrived at this combination as a balance between a strict enough matching requirement to ensure a high degree of similarity between the matched waveforms, and not being too restrictive and eliminating more *in situ* plots from the analysis. Tests using at least 50, 100, 200 and 500 did not yield very large differences in AGBD prediction or uncertainty, and 100 was chosen to include enough FF matches to ensure a robust AGBD distribution, while

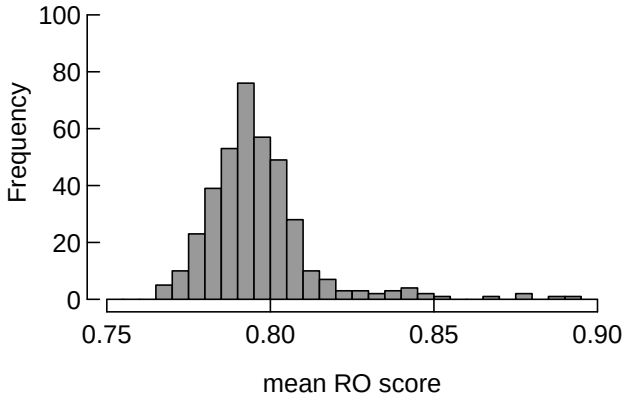

Figure S2: FF produces many stands with waveforms that match the stem mapped field plots well. This distribution shows the mean relative overlap score across the 380 sites with at least 100 matches with a relative overlap above 0.75

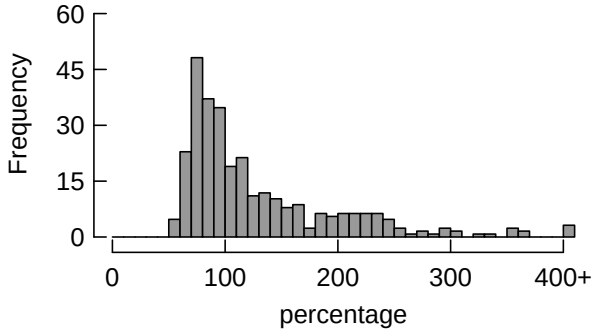

Figure S3: WFM AGBD prediction uncertainty as a percentage of the predicted AGBD value is shown for the 380 *in situ* plots with at least 100 best matches; the uncertainty tends to be similar in magnitude to the predicted value, although some sites have uncertainties several times the predicted AGBD.

keeping the number of matches relatively small to ensure only the waveforms with the highest degree of similarity were used.

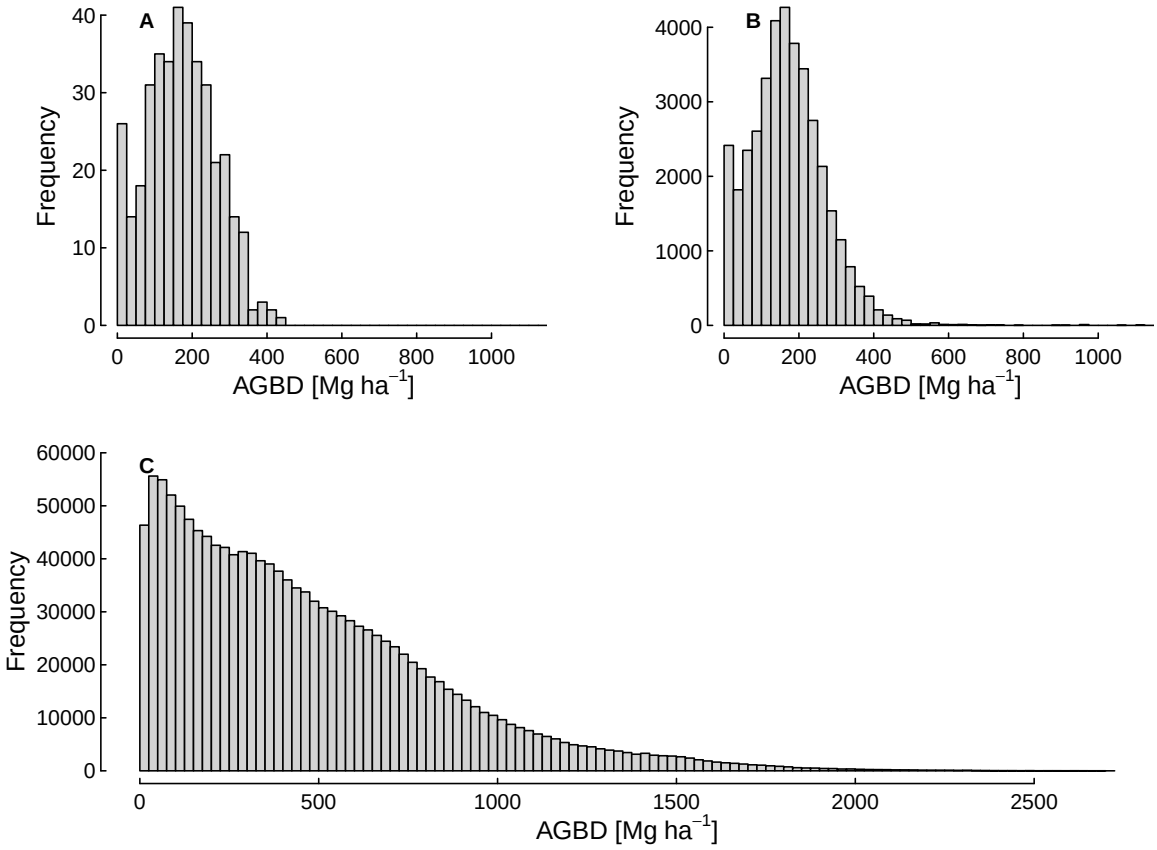

Figure S4: The distribution of AGBD across the 380 *in situ* stem maps with at least 100 best matches (A) is highly similar to the distribution of AGBD across all the FF stands identified as best matches (B), despite a very different distribution of AGBD across the entire FF database used in the WFM process (C).

5. Stem maps and biomass

The 380 *in situ* stem maps with at least 100 best matches had AGBD values that ranged from 2.2 Mg ha^{-1} to 436.1 Mg ha^{-1}, with a mean and median of 171.2 Mg ha^{-1} and 169.7 Mg ha^{-1} respectively (figure S4A). The database of FF stands had AGBD values that ranged from 0.01 Mg ha^{-1} to 2713.6 Mg ha^{-1}, with a mean and median of 469.0 Mg ha^{-1} and 387.8 Mg ha^{-1} (figure S4C). As a whole, the 38,000 forest stands that were identified as best matches (100 matches across 380 sites) had biomass values that ranged from 0.3 Mg ha^{-1} to 1103.6 Mg ha^{-1}, with a mean and median of 170.2 Mg ha^{-1} and 165.1 Mg ha^{-1} respectively (figure S4B).

References

[1] Stanke H, Finley A O, Weed A S, Walters B F and Domke G M 2020 *Environmental Modelling & Software* **127** 104664

[2] Iverson L R, Prasad A M, Matthews S N and Peters M 2008 *Forest ecology and management* **254** 390–406

[3] Jenkins J C, Chojnacky D C, Heath L S and Birdsey R A 2003 *Forest science* **49** 12–35

[4] Bohn F J and Huth A 2017 *Royal Society open science* **4** 160521

[5] Knapp N, Huth A and Fischer R 2021 *Remote Sensing* **13** 1592

[6] Fischer R, Bohn F, de Paula M D, Dislich C, Groeneveld J, Gutiérrez A G, Kazmierczak M, Knapp N, Lehmann S, Paulick S *et al.* 2016 *Ecological Modelling* **326** 124–133

[7] Duncanson L, Neuenschwander A, Hancock S, Thomas N, Fatoyinbo T, Simard M, Silva C A, Armston J, Luthcke S B, Hofton M *et al.* 2020 *Remote Sensing of Environment* **242** 111779

[8] Duncanson L, Kellner J R, Armston J, Dubayah R, Minor D M, Hancock S, Healey S P, Patterson P L, Saarela S, Marselis S, Silva C E, Bruening J, Goetz S J, Tang H, Hofton M, Blair B, Luthcke S, Fatoyinbo L *et al.* In review *Remote Sensing of Environment*